软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(4):1051−1062 [doi: 10.13328/j.cnki.jos.005926]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

梯度有偏情形非光滑问题 NAG 的个体收敛性*

刘宇翔, 程禹嘉, 陶 卿

(中国人民解放军陆军炮兵防空兵学院 信息工程系,安徽 合肥 230031)

通讯作者: 陶卿, E-mail: qing.tao@ia.ac.cn

摘 要: 随机优化方法已经成为处理大规模正则化和深度学习优化问题的首选方法,其收敛速率的获得通常都建立在目标函数梯度无偏估计的基础上,但对机器学习问题来说,很多现象都导致了梯度有偏情况的出现.与梯度无偏情形不同的是,著名的 Nesterov 加速算法 NAG(Nesterov accelerated gradient)会逐步累积每次迭代中的梯度偏差,从而导致不能获得最优的收敛速率甚至收敛性都无法保证.近期的研究结果表明,NAG 方法也是求解非光滑问题投影次梯度关于个体收敛的加速算法,但次梯度有偏对其影响的研究未见报道.针对非光滑优化问题,证明了在次梯度偏差有界的情况下,NAG 能够获得稳定的个体收敛界,而当次梯度偏差按照一定速率衰减时,NAG 仍然可获得最优的个体收敛速率.作为应用,得到了一种无需精确计算投影的投影次梯度方法,可以在保持收敛性的同时较快地达到稳定学习的精度.实验验证了理论分析的正确性及非精确方法的性能.

关键词: 机器学习;Nesterov 加速方法;随机优化;梯度估计有偏;个体收敛

中图法分类号: TP181

中文引用格式: 刘宇翔,程禹嘉,陶卿.梯度有偏情形非光滑问题 NAG 的个体收敛性.软件学报,2020,31(4):1051−1062. http://www.jos.org.cn/1000-9825/5926.htm

英文引用格式: Liu XY, Cheng YJ, Tao Q. Individual convergence of NAG with biased gradient in nonsmooth cases. Ruan Jian Xue Bao/Journal of Software, 2020,31(4):1051−1062 (in Chinese). http://www.jos.org.cn/1000-9825/5926.htm

Individual Convergence of NAG with Biased Gradient in Nonsmooth Cases

LIU Yu-Xiang, CHENG Yu-Jia, TAO Qing

(Department of Information Engineering, PLA Army Academy of Artillery and Air Defense, Hefei 230031, China)

Abstract: Stochastic method has become the first choice for dealing with large-scale regularization and deep learning optimization problems. The acquisition of its convergence rate heavily depends on the unbiased gradient of objective functions. However, for machine learning problems, many scenarios can result in the appearance of biased gradient. In contrast to the unbiased gradient cases, the well-known Nesterov accelerated gradient (NAG) accumulates the error caused by the bias with the iteration. As a result, the optimal convergence will no longer hold and even the convergence cannot be guaranteed. Recent research shows that NAG is also an accelerated algorithm for the individual convergence of projection sub-gradient methods in non-smooth cases. However, until now, there is no report about the affect when the subgradient becomes biased. In this study, for non-smooth optimization problems, it is proved that NAG can obtain a stable individual convergence bound when the subgradient bias is bounded, and the optimal individual convergence can still be achieved while the subgradient errors decrease at an appropriate. As an application, an inexact projection subgradient method is obtained in which the projection needs not calculate accurately. The derived algorithm can approach the stable learning accuracy more quick while keeping the convergence. The experiments verify the correctness of theoretical analysis and the performance of inexact methods.

Key words: machine learning; Nesterov accelerated gradient; stochastic optimization; biased gradient; individual convergence

* 基金项目: 国家自然科学基金(61673394)

Foundation item: National Natural Science Foundation of China (61673394)

本文由“非经典条件下的机器学习方法”专题特约编辑高新波教授、黎铭教授、李天瑞教授推荐.

收稿时间: 2019-05-31; 修改时间: 2019-08-01; 采用时间: 2019-09-20; jos 在线出版时间: 2020-01-10

CNKI 网络优先出版: 2020-01-14 09:53:25, http://kns.cnki.net/kcms/detail/11.2560.TP.20200114.0953.011.html

顾名思义,随机优化方法是经典梯度优化方法的随机形式,其每一步迭代仅需使用目标函数梯度的无偏估计,主要用于求解目标函数是期望形式的优化问题[1].对于机器学习问题,当样本点独立同分布时,单个样本所对应的目标函数梯度正好是整个目标函数梯度的无偏估计[2],随机优化算法从而避免了每一步迭代都遍历样本集计算梯度的问题.同时,大规模样本集合往往存在着冗余现象,因此只要针对部分样本运行随机优化方法后,就能达到稳定的学习精度[1].正是由于随机优化方法具有这些特点,很多经典的梯度优化方法被推广成随机形式并取得了最优的收敛速率[3,4],但它们都是限制在梯度估计是无偏的这一条件下[5].

在机器学习优化问题的求解中,梯度估计有偏现象是经常出现的.首先,样本集是否独立来源于未知分布,并不知晓,很难保证随机优化算法抽取样本所对应的目标函数的梯度正好是整个目标函数梯度的无偏估计[6].其次,机器学习广泛使用的 Proximal 优化方法的每一步迭代都涉及到优化子问题的求解,一些类型的子问题只能近似求解,这种求解方式可以归结到梯度估计有偏情形下的精确求解[7].第三,对于光滑优化问题,目标函数的 Lipschitz 常数通常决定了算法的步长,这个常数估计中的误差问题也可以归结为一种梯度有偏估计问题[8].

NAG(Nesterov accelerated gradient)是求解光滑问题占统治地位的一种优化方法,将梯度下降算法的收敛速率从 $O(1/t)$加速至最优的 $O(1/t^2)$[9],其中,t 为迭代次数.它填补了 Nemirovski 与 Yudin 所证明的“任何一阶优化方法都不可能取得比 $O(1/t^2)$更快的收敛速率”的空白[10].但当梯度估计有偏时,Devolder 和 Nesterov 等人分析指出,随机 NAG 方法每次迭代中梯度有偏导致的偏差项可能会随着迭代次数的继续呈线性增长[11],从而导致不能获得应有的收敛速率甚至收敛性都无法保证.相比之下,随机梯度下降法每次迭代产生的偏差就不会累积,最终偏差项仍为一个常数.这意味着加速算法在给我们带来收敛速率数量级式加速的同时,其对梯度计算的要求也相当苛刻,而这一问题却被很多应用人员所忽视.实际上,只有在梯度的偏差满足一定衰减速率的条件下,NAG 方法才能保持原有的收敛速率[9].梯度估计有偏对其他类型的加速算法也导致了同样的问题,也引起了广大学者们的关注[2,7,12],如 2018 年 Julian 等人对一阶 Primal-dual 优化方法即进行了研究[13].

对于非光滑优化问题,投影次梯度方法关于平均输出可以达到最优的收敛速率($O(1/\sqrt{t})$),但个体形式输出仅能获得 $O(\log t/\sqrt{t})$ 的收敛速率[14].通过选取合适的步长策略,NAG 可以将投影次梯度方法的个体收敛速率加速至最优的 $O(1/\sqrt{t})$[15].与平均输出方式的解相比,NAG 的个体解在求解稀疏学习问题应用中可以获得更好的稀疏性.由此产生一个不能回避的问题,即非光滑问题的个体收敛加速算法 NAG 是否也受到次梯度估计偏差带来的影响?此时如何确保其最优收敛速率?

本文主要研究梯度有偏情形下,求解非光滑优化问题个体收敛加速算法随机 NAG 的收敛性问题,主要贡献如下.

(1) 获得了非光滑问题个体收敛加速算法 NAG 基于梯度偏差的个体收敛界.由于非光滑函数与光滑函数的性质不同,导致光滑 NAG 收敛界的证明思路[7]在这里并不适用.我们的思路是将次梯度偏差项引入到非光滑情形 NAG 个体收敛性的证明中[15],通过选取合适的步长,并对次梯度偏差导致的项作适当的处理.其次,在实际应用中,梯度偏差是否满足相应的衰减性质很难验证.然而,在非光滑情形下,NAG[15]由于采用变步长策略,获得的收敛速率比光滑情形要慢,但受偏差项的影响要弱一些,仅需次梯度偏差有界就可获取收敛界,这是在光滑有偏情形下 NAG[7]无法取得的结论.

(2) 得到了一种无需精确计算投影的投影次梯度方法,可以在保持收敛性的同时较快地达到稳定的学习精度.投影计算是投影次梯度方法难以克服的计算瓶颈,在实际运行中会导致循环嵌套循环的问题.为了解决这一问题,一些文献采用了不使用投影或者仅在最后一步迭代中使用投影而中间过程不进行投影运算的思路[16,17].显然,投影次梯度的每一步迭代等价于解析求解一个优化子问题,因此使用不精确的投影运算可以视为次梯度有偏的问题,本文的理论分析确保这种不精确投影次梯度方法的收敛界,通过减少嵌套循环次数较快地达到了稳定学习的精度.

(3) 实验验证了理论分析的正确性及非精确方法的性能.在人工数据集上通过选取不同的偏差衰减参数验证了理论分析的正确性;在标准数据集上通过与精确投影方法相比,验证了非精确投影方法能够快速达到稳定的学习精度.

1 NAG方法与梯度有偏问题

以二分类为例,假设训练样本集 $S=\{\xi_i=(\boldsymbol{x}_i,y_i)|i=1,2,...,m\}\subseteq R^n\times\{+1,-1\}$ 满足独立同分布.本文仅考虑如下典型的非光滑正则化学习问题:

$$\min F(\boldsymbol{w})=r(\boldsymbol{w})+\frac{1}{m}\sum_{i=1}^{m}f_i(\boldsymbol{w}) \tag{1}$$

其中,$r(\boldsymbol{w})=\lambda\|\boldsymbol{w}\|_1$ 为 l_1 正则化项,$f_i(\boldsymbol{w})$为单个样本 ξ_i 对应的 hinge 损失函数.对于给定的参数 z,问题(1)等价于如下约束问题[18]:

$$\begin{cases}\min f(\boldsymbol{w})=\dfrac{1}{m}\sum_{i=1}^{m}f_i(\boldsymbol{w}),\\ \text{s.t.}\|\boldsymbol{w}\|_1\leqslant z\end{cases} \tag{2}$$

对于求解上述问题,投影次梯度方法批处理形式的关键迭代步骤为[19]

$$\boldsymbol{w}_{t+1}=\boldsymbol{P}_Q(\boldsymbol{w}_t-\eta_t\boldsymbol{g}(\boldsymbol{w}_t)) \tag{3}$$

其中,闭凸集 Q 为 l_1 范数球 $\{\boldsymbol{w}|\|\boldsymbol{w}\|_1\leqslant z\}$, $\boldsymbol{P}_Q$ 为投影算子,$\boldsymbol{g}(\boldsymbol{w}_t)$是 $f(\boldsymbol{w})$在 $\boldsymbol{w}_t$ 处的次梯度.当样本集满足独立同分布时,$\boldsymbol{g}_i(\boldsymbol{w}_t)$是 $\boldsymbol{g}(\boldsymbol{w}_t)$的无偏估计.随机形式投影次梯度方法的关键迭代步骤为

$$\boldsymbol{w}_{t+1}=\boldsymbol{P}_Q(\boldsymbol{w}_t-\eta_t\boldsymbol{g}_{i_t}(\boldsymbol{w}_t)) \tag{4}$$

其中,i_t是指第 t 次迭代中抽取的单个样本ξ_i的序号.通过式(3)与式(4)的比较可以看出,随机算法克服了批处理方法的低效行为[1].

当目标函数 f 为光滑函数,即满足 L-Lipschitz 光滑性质时:

$$f(\boldsymbol{w})\leqslant f(\boldsymbol{u})+\langle\nabla f(\boldsymbol{u}),\boldsymbol{w}-\boldsymbol{u}\rangle+\frac{L}{2}\|\boldsymbol{w}-\boldsymbol{u}\|^2,\forall\boldsymbol{w},\boldsymbol{u}\in R^N \tag{5}$$

传统的投影次梯度方法能够取得 $O(L/t)$的收敛速率[20],其中,$\nabla f(\boldsymbol{u})$为 f 在 $\boldsymbol{u}$ 处的梯度值.NAG 加速算法采用一定的特殊步长策略,利用动量加速技巧结合光滑函数性质构造出递推关系,从而获得了 $O(L/t^2)$的个体收敛速率[9],可表述为如下形式:

$$\begin{cases}\boldsymbol{y}_t=\boldsymbol{w}_t+\theta_t(\theta_{t-1}^{-1}-1)(\boldsymbol{w}_t-\boldsymbol{w}_{t-1})\\ \boldsymbol{w}_{t+1}=\boldsymbol{P}_Q(\boldsymbol{y}_t-L\nabla f(\boldsymbol{y}_t))\\ \theta_{t+1}=()\sqrt{\theta_t^4+4\theta_t^2}-\theta_t^2\big/2\end{cases} \tag{6}$$

类似于式(4),NAG(6)也可推广为随机形式.但在梯度估计有偏的情况下,传统的优化方法(3)每次迭代产生的偏差不会累积,最终偏差项为一个常数δ(收敛界为 $O(L/t+\sigma/\sqrt{t}+\delta)$, 其中,$\sigma$为梯度估计的方差,$\delta$为梯度估计偏差所导致的计算偏差);NAG(6)每次迭代产生的偏差可能会随着迭代次数 t 逐渐累积(收敛界为 $O(L/t^2+\sigma/\sqrt{t}+t\delta)$)[11].

实际上,很多情况都可以归结为梯度有偏情形,典型情况有[7,8,12]:

(1) 非精确方法在每一步迭代都涉及到优化子问题的求解,fused-LASSO 和 nuclear-norm 等类型的子问题只能够采用近似梯度 $\boldsymbol{g}_{prox}$ 计算快速求解梯度值,但与真实梯度值$\nabla f(\boldsymbol{w}_t)$存有偏差$\boldsymbol{\varepsilon}_t=\boldsymbol{g}_{prox}-\nabla f(\boldsymbol{w}_t)$;

(2) 样本集并未满足独立同分布这个假定条件,导致随机抽取单个样本计算的梯度$\nabla f_i(\boldsymbol{w}_t)$与真实梯度值$\nabla f(\boldsymbol{w}_t)$存有偏差$\boldsymbol{\varepsilon}_t=\nabla f_i(\boldsymbol{w}_t)-\nabla f(\boldsymbol{w}_t)$;

(3) 光滑函数通常采用光滑系数作为常步长,估计值 $\bar{L}$ 与光滑系数存有一定的偏差 $\Delta=L-\bar{L}$, 进而导致每次迭代存有相对梯度偏差 $\boldsymbol{\varepsilon}_t=(\bar{L}-L)\nabla f(\boldsymbol{w}_t)/L$.

针对梯度有偏情形,Schmidt 等人考虑了梯度偏差衰减的可能性,并进一步挖掘了梯度计算误差(error)与相应的函数偏差项(bias)之间的潜在关系,证明了随机形式的 NAG(6)在梯度有偏情形下具有如下收敛界[7]:

$$f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)\leqslant\frac{2L}{(t+1)^2}(R+2A_t+\sqrt{2B_t})^2 \tag{7}$$

其中, $A_t=\sum_{i=1}^{t} i\left(\frac{\|e_i\|}{L}+\sqrt{\frac{2\varepsilon_i}{L}}\right), B_t=\sum_{i=1}^{t}\frac{i^2\varepsilon_i}{L}$, e_i 为每次迭代中计算梯度的误差,ε_i 为每次近似梯度计算导致的函数偏差,R 为初始点与最优点的欧式距离.文献[7]指出,当梯度误差 $\|e_i\|$ 和计算偏差 $\sqrt{\varepsilon_i}$ 均保证一定的衰减速率时($O(1/t^{2+a}), a>0$),NAG才能保持原有的$O(1/t^2)$收敛速率;当偏差衰减速率为$O(1/t)$时,NAG才能获取一个收敛界.文献[8]中的实验也进一步验证了 NAG 算法在梯度估计有偏情形下的不稳定性.

文献[15]将 NAG 加速算法推广至非光滑情形:

$$\begin{cases} \boldsymbol{y}_t=\boldsymbol{w}_t+\theta_t(\theta_{t-1}^{-1}-1)(\boldsymbol{w}_t-\boldsymbol{w}_{t-1}) \\ \boldsymbol{w}_{t+1}=\boldsymbol{P}_Q(\boldsymbol{y}_t-\eta_t\boldsymbol{g}(\boldsymbol{y}_t)) \end{cases} \tag{8}$$

证明其取得了 $O(1/\sqrt{t})$ 的最优个体收敛速率,并具有良好的稀疏性.不同于光滑情形 NAG(6),NAG(8)采用的是变步长,而 NAG(6)选取的步长是常数.在梯度有偏情形下,NAG(6)的收敛界已在文献[7]中给出讨论,下面我们将讨论 NAG(8)的收敛界.

2 梯度有偏情形非光滑 NAG 的个体收敛性分析及其应用

本节主要在 Schmidt 等人的基础上进一步提出了梯度有偏问题更为一般情形的定义,证明分析在偏差有界的情况下,NAG 在非光滑次梯度估计有偏情形时的收敛界以及其与次梯度偏差之间的关系.同时,为了降低投影问题的求解时间开销,提出了一种无需精确计算投影的 I-NAG(inexact Nesterov accelerated gradient)方法.

2.1 NAG 方法的个体收敛界分析

本文我们沿用文献[12]的思想,将非光滑情形下次梯度有偏情形统一归结为第 t 次迭代运算计算次梯度值 $\hat{\boldsymbol{g}}(\boldsymbol{w}_t)$ 与真实次梯度值 $\boldsymbol{g}(\boldsymbol{w}_t)$存有一个偏差$\boldsymbol{\varepsilon}_t$:

$$\boldsymbol{g}(\boldsymbol{w}_t)=\hat{\boldsymbol{g}}(\boldsymbol{w}_t)+\boldsymbol{\varepsilon}_t \tag{9}$$

hinge 损失是机器学习中最典型的非光滑损失函数,其在次梯度无偏估计时具有如下性质:

$$f(\boldsymbol{w}_{t+1})\leqslant f(\boldsymbol{w}_t)+\langle \boldsymbol{g}(\boldsymbol{w}_t),\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\rangle+M\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\| \tag{10}$$

其中,M 为 hinge 损失函数的 Lipschitz 常数.进一步考虑次梯度估计有偏的情形,有:

$$f(\boldsymbol{w}_{t+1})\leqslant f(\boldsymbol{w}_t)+\langle \hat{\boldsymbol{g}}(\boldsymbol{w}_t)+\boldsymbol{\varepsilon}_t,\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\rangle+M\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\| \tag{11}$$

为了方便讨论、分析,我们将 NAG(8)写成如下等价形式:

$$\begin{cases} \boldsymbol{y}_t=\boldsymbol{w}_t+\theta_t(\theta_{t-1}^{-1}-1)(\boldsymbol{w}_t-\boldsymbol{w}_{t-1}) \\ \boldsymbol{w}_{t+1}=\arg\min\limits_{\boldsymbol{w}\in Q}\left\{\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{y}_t\|^2+\eta_t\langle \boldsymbol{g}(\boldsymbol{y}_t),\boldsymbol{w}-\boldsymbol{y}_t\rangle\right\} \end{cases} \tag{12}$$

在讨论、分析 NAG(8)在非光滑条件下的收敛速率问题之前,首先介绍一下本文证明的关键引理.

引理 1(三点性质). 假设 f 为满足下半连续的凸函数,$\boldsymbol{g}(\boldsymbol{w}_t)$为 f 在 $\boldsymbol{w}_t$ 时的次梯度,且 $\boldsymbol{w}_{t+1}$ 满足以下式子:

$$\boldsymbol{w}_{t+1}=\arg\min_{\boldsymbol{w}}\left\{\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{w}_t\|^2+\eta_t\langle \boldsymbol{g}(\boldsymbol{w}_t),\boldsymbol{w}-\boldsymbol{w}_t\rangle\right\} \tag{13}$$

则有以下不等式成立:

$$\frac{1}{2}\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|^2+\eta_t\langle \boldsymbol{g}(\boldsymbol{w}_t),\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\rangle+\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{w}_{t+1}\|^2\leqslant\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{w}_t\|^2+\eta_t\langle \boldsymbol{g}(\boldsymbol{w}_t),\boldsymbol{w}-\boldsymbol{w}_t\rangle \tag{14}$$

其相关详细内容及证明请参见文献[21]中的引理 3.2.

引理 2(投影区域的有界性). 假设 Q 为有界闭凸集,则$\exists D>0$,使得$\forall \boldsymbol{x},\boldsymbol{y}\in Q$,均有 $\|\boldsymbol{x}-\boldsymbol{y}\|\leqslant D$.

引理 2 的相关证明可参见文献[19].依据引理 1 和引理 2,参照 NAG 在次梯度无偏情形下的证明思路[15],我们可以得出以下定理成立,相关证明详见附录 1.

定理 1. 假设 Q 为 R^N 空间上的有界闭凸集,f 为上满足下半连续的一般凸函数,M 为 hinge 损失函数的

Lipschitz 常数,初始点 $\boldsymbol{w}_1 \in R^N$, $\{\boldsymbol{w}_t\}_{t=1}^{\infty}$ 由式(8)产生,则$\exists D>0$,使得$\forall \boldsymbol{w} \in Q$ 有:

$$f(\boldsymbol{w}_{t+1}) - f(\boldsymbol{w}) \leqslant \frac{\theta_t^2}{2\eta_t} D^2 + \theta_t^2 \sum_{k=0}^{t} \frac{\eta_k}{2\theta_k^2} M^2 + \theta_t^2 \sum_{k=0}^{t} \frac{1}{\theta_k} \|\boldsymbol{\varepsilon}_k\| D \tag{15}$$

进一步地,我们将 NAG(8)推广至随机形式:

$$\begin{cases} \boldsymbol{y}_t = \boldsymbol{w}_t + \theta_t(\theta_{t-1}^{-1} - 1)(\boldsymbol{w}_t - \boldsymbol{w}_{t-1}) \\ \boldsymbol{w}_{t+1} = \boldsymbol{P}_Q(\boldsymbol{y}_t - \eta_t \boldsymbol{g}_{i_t}(\boldsymbol{y}_t)) \end{cases} \tag{16}$$

文献[22]中的引理 1 给出了一种将批处理算法的收敛速率转化为随机形式的证明技巧,我们可以采用类似的方法将定理 1 推广至随机形式,得到定理 2.

定理 2. 假设 Q 为 R^N 空间上的有界闭凸集,f 为 Q 上满足下半连续的一般凸函数,M 为 hinge 损失函数的 Lipschitz 常数,则$\exists D>0$,使得$\forall \boldsymbol{w} \in Q$ 有:

$$E[f(\boldsymbol{w}_{t+1}) - f(\boldsymbol{w})] \leqslant \frac{\theta_t^2}{2\eta_t} D^2 + \theta_t^2 \sum_{k=0}^{t} \frac{\eta_k}{2\theta_k^2} M^2 + \theta_t^2 \sum_{k=0}^{t} \frac{1}{\theta_k} E\left[\|\boldsymbol{\varepsilon}_k\|\right] D \tag{17}$$

通过对比文献[15]中定理 3 不难看出,在梯度估计有偏的情形下,函数的相对收敛界多了 $\theta_t^2 \sum_{k=0}^{t} \frac{1}{\theta_k} E\left[\|\boldsymbol{\varepsilon}_k\|\right] D$ 这一偏差项.当取步长 $\theta_t = 1/(t+2), \eta_t = 1/(t+2)\sqrt{t+2}$ 时,上面不等式右边前两项之和的数量级为 $O(1/\sqrt{t})$,然而,$\theta_t^2 \sum_{k=0}^{t} \frac{1}{\theta_k}$ 为一个常数,故而 $\{\boldsymbol{\varepsilon}_t\}_{t=1}^{\infty}$ 的性质直接影响到目标函数的收敛速率.从而我们对梯度误差 $\{\|\boldsymbol{\varepsilon}_t\|\}_{t=1}^{\infty}$ 的相关性质进行假定,并假设 $\boldsymbol{w}^*$为理论上的最优点,不难得出推论 1 和推论 2 的结论.

推论 1. 假设偏差 $\{\|\boldsymbol{\varepsilon}_t\|\}_{t=1}^{\infty}$ 有界,$\exists D, \delta > 0$,使得 $\max_{\varepsilon} \|\boldsymbol{\varepsilon}_t\| \leqslant \delta / D$,$f$ 为有界闭凸集 $Q \subseteq R^N$ 上满足下半连续的一般凸函数,当取步长 $\theta_t = 1/(t+2), \eta_t = 1/(t+2)\sqrt{t+2}$,$\forall \boldsymbol{w} \in Q$ 时,则有:

$$E[f(\boldsymbol{w}_{t+1}) - f(\boldsymbol{w}^*)] \leqslant \frac{D^2 + M^2}{2\sqrt{t+2}} + \delta \tag{18}$$

推论 2. 假设偏差 $\{\|\boldsymbol{\varepsilon}_t\|\}_{t=1}^{\infty}$ 随着迭代次数逐渐衰减($O(1/t^{1/2+a}), a>0$),$\exists D, \delta > 0$,使得 $\max_{\varepsilon} \|\boldsymbol{\varepsilon}_t\| \leqslant \delta / D$,$f$ 为有界闭凸集 $Q \subseteq R^N$ 上满足下半连续的一般凸函数,当取步长 $\theta_t = 1/(t+2), \eta_t = 1/(t+2)\sqrt{t+2}$,$\forall \boldsymbol{w} \in Q$ 时,则有:

$$E[f(\boldsymbol{w}_{t+1}) - f(\boldsymbol{w}^*)] \leqslant \frac{D^2 + M^2}{2\sqrt{t+2}} + \frac{1}{t^{\frac{1}{2}+a}} \delta \tag{19}$$

推论 1 和推论 2 更为直观地描绘出在非光滑情形下随机形式的 NAG(8)的个体收敛速率与次梯度偏差的关系,即当次梯度偏差 $\{\|\boldsymbol{\varepsilon}_t\|\}_{t=1}^{\infty}$ 有界时,NAG(8)在每次迭代中所造成的偏差能够累积为一个常数;当次梯度偏差 $\{\|\boldsymbol{\varepsilon}_t\|\}_{t=1}^{\infty}$ 呈一定速率衰减时($O(1/t^{1/2+a}), a>0$),NAG(8)能够获取在次梯度估计无偏时同等的收敛速率.

至此,我们将光滑有偏情形下的 NAG(6)的收敛界研究推广至非光滑情形.与 Schmidt 等人研究内容[7]的差异在于,我们考虑了非光滑情形与光滑情形的函数性质区别,采用了不同的证明方式获取了 NAG(8)的个体收敛界.在实际应用中,梯度偏差是否满足对应的衰减速率较难验证,然而,NAG(8)采用变步长策略,因而受梯度偏差的影响较弱,仅需次梯度偏差有界即可获取收敛界;然而,在光滑有偏情形下,NAG(6)仍需梯度偏差衰减速率为 $O(1/t)$才能获取收敛界.

2.2 一种非精确投影形式的 NAG 方法

许多随机梯度下降方法都需要在每次迭代时进行投影运算,以确保所获得的解保持在定义域内.精确投影运算通常采用二分法来寻找可行解,单次运算最多需要 $\log_2\left(\frac{\beta-\alpha}{\gamma}\right)$ 次内循环运算[18],其中,α和β为初始的可行解区间$[\alpha,\beta]$的上下界,γ为设定的投影精度参数.投影计算也是投影次梯度方法难以克服的计算瓶颈,在实际运行中会导致出现循环嵌套循环的问题,时间消耗较大[17].为了解决这一问题,我们借鉴一些文献采用不使用投影

或仅在最后一步使用投影的思路[16,17],考虑将每次投影运算的精度降低(即增大投影精度参数γ的值),进而有效减少了嵌套循环的次数,提出了一种非精确投影的随机 I-NAG 方法,其关键迭代步骤为

$$\begin{cases} \boldsymbol{y}_t = \boldsymbol{w}_t + \theta_t(\theta_{t-1}^{-1}-1)(\boldsymbol{w}_t - \boldsymbol{w}_{t-1}) \\ \boldsymbol{w}_{t+1} = \tilde{\boldsymbol{P}}_Q(\boldsymbol{y}_t - \eta_t \boldsymbol{g}_{i_t}(\boldsymbol{y}_t)) \end{cases} \tag{20}$$

其中, $\tilde{\boldsymbol{P}}_Q$ 为 Q 上非精确形式的投影算子,采用低精度的 $\tilde{\gamma}$ 值进行投影运算,其主要执行过程见算法 1.

算法 1. I-NAG 方法.

1: Initialize a weight vector $\boldsymbol{w}_1$
2: **For** t =1 to T
3: Compute $\theta_t = 1/(t+2)$, $\boldsymbol{y}_t$ via (20).
4: Compute $\eta_t = 1/(t+2)\sqrt{t+2}$, $\boldsymbol{g}_{i_t}(\boldsymbol{y}_t)$ via (20).
5: Compute $\boldsymbol{w}_{t+1}$ via (20).
6: **End for**
7: Output: $\boldsymbol{w}_{T+1}$

非精确方法每一步迭代都涉及到优化子问题的求解,一些类型的子问题只能近似求解,Schmidt 等人将这种求解方式归结到梯度估计有偏情形下的精确求解.显然,投影次梯度的每一步迭代等价于解析求解一个优化子问题,非精确投影方法通过降低每次迭代投影运算的精度,也可以看作是一种类似的非精确方法,因此非精确投影算法导致的梯度偏差也可视为一种梯度估计有偏情形.

3 数值实验

本节通过 1 个人工数据集和 4 个标准大规模数据集验证 NAG 方法在次梯度估计有偏情形下的理论分析和检验 I-NAG 方法的性能.

本文实验参照文献[23]的方法求解 hinge 损失函数的次梯度,采用文献[18]中精确求解 l_1 投影球的方法求解投影问题.程序中投影运算采用刘军的 SLEP 投影工具箱[24]的函数 eplb,其原理是通过二分法求得精确解.为了公平起见,算法每次迭代过程中所有样本采取随机抽取的形式,并重复 10 次实验后取结果的算术平均值作为最后输出进行比较,且对同一数据集程序中各算法都采用相同的参数.

3.1 人工数据集的实验结果及分析

实验采用了文献[25]的方法构建人工数据集,其中训练样本个数为 10 000,测试样本个数为 2 000,样本维数均为 100.其中,向量各分量值均服从正态分布 $N(0,1)$,并随机从中抽取 10 个数替换为 0.同时生成随机噪声(分量服从正态分布 $N(0,1)$),并令每个样本 i 的标签为 $y_i=X_i\boldsymbol{w}+\sigma_i$($X$ 为训练集或测试集样本矩阵,σ_i 为第 i 个样本对应的噪声),当 $y_i>0$ 时,令 $y_i=1$;当 $y_i<0$ 时,令 $y_i=-1$.由此建立了人工有偏样本数据集.

本文参照文献[7]中的方法进行对比实验,即首先生成随机偏差$\boldsymbol{\varepsilon}\sim N(0,1)$,对其增加不同程度衰减系数 a(a=0,0.2,0.6,+∞),形成新的偏差$\boldsymbol{\varepsilon}_t=\boldsymbol{\varepsilon}_t/t^a$,而后加入 NAG[16]每次迭代的次梯度值中,其收敛性情况如图 1 所示.

图 1 中横坐标表示算法迭代次数;纵坐标表示为相对目标函数值,即当前目标函数值与最优目标函数值之差的对数值 $\log(f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*))$,$f(\boldsymbol{w}^*)$取迭代中最小的目标函数值.图中 4 种曲线分别表示为:绿色点虚线表示次梯度偏差未衰减(a=0);紫色点线表示次梯度偏差呈 $O(1/t^{1/5})$速率衰减(a=0.2);蓝色虚线表示次梯度偏差呈 $O(1/t^{3/5})$速率衰减(a=0.6);红色实线表示次梯度估计无偏下的 NAG 情况(a=+∞).

从图 1 可以看出,当迭代达到一定次数后,a=0 时 NAG 收敛到一定的数值;a=0.2 时收敛十分缓慢,原因在于,当次梯度偏差衰减速率小于 $O(1/t^{1/2})$时,偏差项影响了收敛速率,从而与推论 1 的结论相吻合;当 a=0.6 和 a=+∞时均保持同等的收敛速率,原因在于,当偏差项收敛速率大于等于 $O(1/t^{1/2})$时,偏差项的累积不影响算法的收敛速率,从而与推论 2 的结论相吻合.

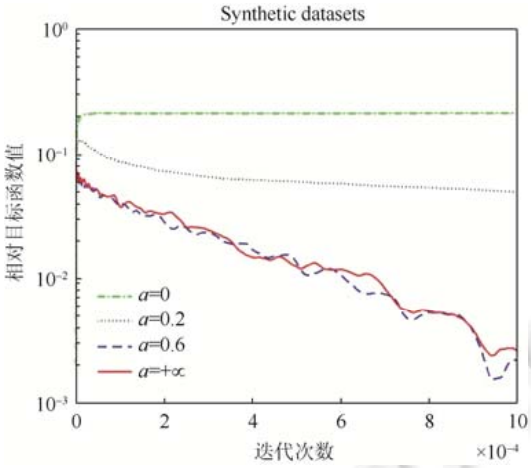

Fig.1 The influence on the convergence rate when using different a

图 1 a 的取值对算法收敛速率的影响图

3.2 标准数据集的实验结果及分析

本实验的主要目的是通过与精确投影方法进行对比,验证 I-NAG 的性能.所采用的标准数据集(astro、ijcnn1、covtype 和 rcv1 均来自台湾大学林志仁 Libsvm 网站)相关信息见表 1.

Table 1 Introduction of standard datasets

表 1 标准真实数据集描述

数据集	训练样本	测试样本	维度
astro	29 882	32 487	99 757
ijcnn1	49 990	91 701	22
covtype	522 911	58 101	54
rcv1	20 242	677 399	47 236

为了与原方法进行对比,我们通过将精度评判标准γ由 10^{-8} 降低为 10^{-1}(SLEP 投影工具箱[24]中 eplb 函数 delta 变量),形成非精确投影求解问题的 I-NAG 方法(即本文算法 1).对于非光滑问题,目前获得最优个体收敛速度的优化算法比较少,因此本文仅增添了具有个体最优速度的线性插值投影次梯度方法(projected subgradient method with linear interpolation operation,简称 PSM-I)[26]和 NAG[15]这两种算法作为比较对象.实验中,根据各算法的收敛性结论确定步长设置,其中,NAG 和 I-NAG 均选用同样的步长θ_t=1/(t+2)和 $\eta_t=1/(t+2)\sqrt{t+2}$, PSM-I 的步长为 a_t=1 和 $\eta_t=1/\sqrt{t}$. 3 种方法均通过调用 SLEP 投影工具箱[24]进行投影运算,选取相同的约束参数 z,其中,投影约束域 Q 为 l_1 范数球 $\{\boldsymbol{w}\,|\,\|\boldsymbol{w}\|_1\leqslant z\}$, z 根据数据集的不同选取对应的值(影响算法的收敛性和稀疏性).

在评价优化算法性能时,一般都采用收敛速率作为评价标准,而在评判学习算法性能时,学习精度是最终关注的标准.收敛速率和学习精度固然同等重要,但当目标函数达到一定精度时,花费大量时间优化运算,并不能使学习精度得到显著提升[27].本次实验,我们考虑利用准确率来代替收敛精度作为评价标准.

在实际应用中,我们更关心的是优化算法达到稳定测试精度所需要的时间开销.由于算法之间的主要差别在于投影运算上,投影运算时间也是影响算法训练时间的重要因素.考虑不同的编译环境和设备可能导致 CPU 时间等实验效果有所出入,我们借鉴了文献[28]中以循环次数 epoch 为评价标准的做法,选取投影运算中内循环次数为标准.内循环总次数是指算法在整个训练过程中,投影运算通过二分类方法循环的总次数,是影响整个算法迭代运行时间的主要因素.

图 2 为基于迭代次数的准确率比较图,其中纵坐标表示准确率,即当前输出个体解在测试集正确分类的比例,准确率越高,表示算法的实际分类效果越好.图中蓝色实线表示为非精确方法 I-NAG 的收敛趋势;绿色点虚线表示为 NAG 的收敛趋势;红色虚线表示为线性差值方法 PSM-I 的收敛趋势.从图 2 可以看出:在同等的迭代次数

条件下,在 icjnn1 和 covtype 这两个低维数据集上三者差别不大.然而,PSM-I 为线性差值输出的解,稀疏性较差,且在梯度有偏情形下的收敛性尚未有理论分析.在 astro 和 rcv1 样本维度较高的数据集上,I-NAG 相对另外两种方法能够快速地达到稳定的学习精度,验证了 I-NAG 方法的有效性.

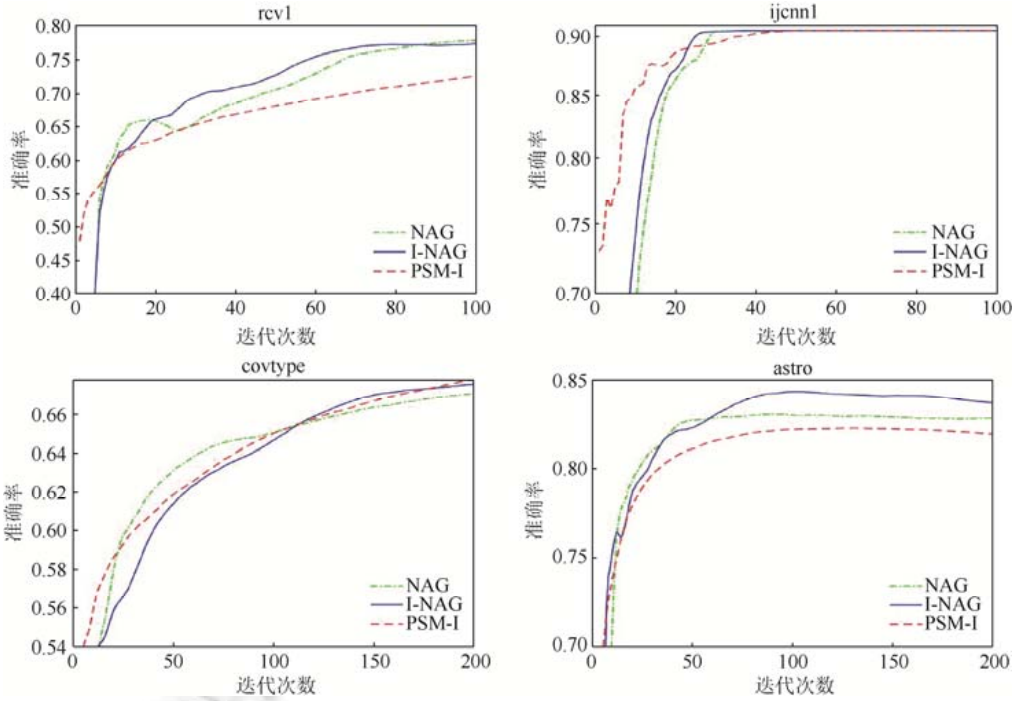

Fig.2 Comparisons of accuracy rate in standard datasets (iterations)

图 2 标准数据集目标函数准确率比较图(迭代次数)

图 3 所示横坐标表示的是图 2 中算法迭代对应的投影内循环总次数,通过与图 2 对比可以更明显地看出差异性.在 astro 和 rcv1 样本维度较高的数据集,I-NAG 仅需相对较少的内循环次数即可获取可观的学习精度.主要原因在于,在高维度的数据集中投影运算相对更为复杂,单次投影所需内循环次数更多,非精确投影能够有效地减少投影内循环次数,进而快速达到稳定的学习精度.综上可以看出,I-NAG 方法在处理高维数据时具有良好的性能.

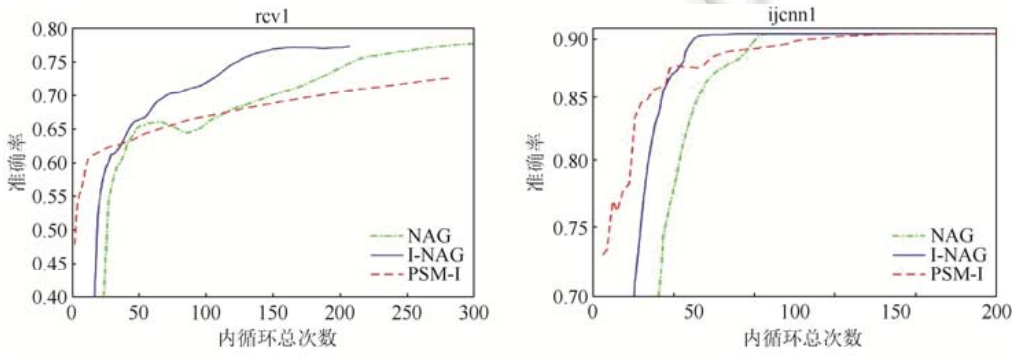

Fig.3 Comparisons of convergence rate in standard datasets (inner iterations)

图 3 标准数据集目标函数收敛速率比较图(内循环总次数)

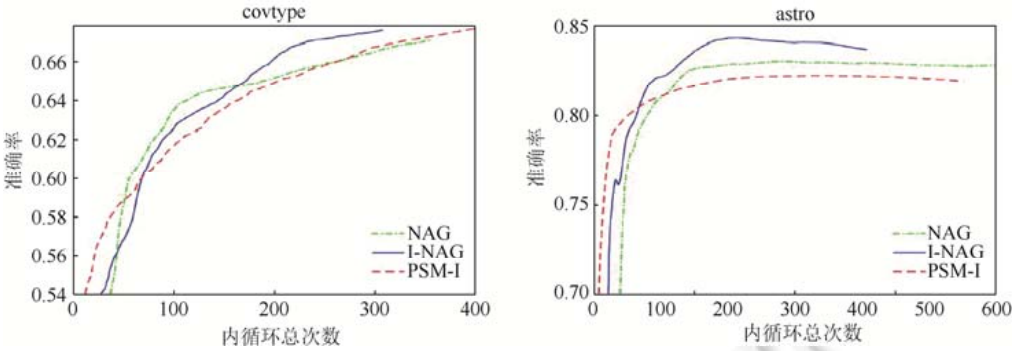

Fig.3 Comparisons of convergence rate in standard datasets (inner iterations) (Continued)

图 3 标准数据集目标函数收敛速率比较图(内循环总次数)(续)

4 总结与展望

本文在非光滑情况下,研究分析了个体收敛加速算法 NAG 基于次梯度偏差的个体收敛界.相对于光滑情形,非光滑情形时 NAG 收敛速率较弱,故而对次梯度估计偏差的敏感性不是很强.作为应用,我们得到了一种无需精确计算投影的 I-NAG 方法,可以通过减少投影内循环次数,在保持收敛性的同时快速达到稳定的学习精度.最后,通过实验验证了理论分析的正确性和所提算法的有效性.下一步我们将研究分析 NAG 在强凸情形下个体收敛性与次梯度偏差的关系,同时还会考虑如何定量化权衡投影子问题的求解精度和学习精度之间的关系.

References:

[1] Bottou L, Curtis FE, Nocedal J. Optimization methods for large-scale machine learning. SIAM Review, 2018,60(2):223−311.

[2] Nemirovski A, Juditsky A, Lan G, Shapiro A. Robust stochastic approximation approach to stochastic programming. SIAM Journal on Optimization, 2009,19(4):1574−1609. [doi: 10.1137/070704277]

[3] Hu C, Pan W, Kwok JT. Accelerated gradient methods for stochastic optimization and online learning. In: Advances in Neural Information Processing Systems. 2009. 781−789.

[4] Xiao L. Dual averaging methods for regularized stochastic learning and online optimization. Journal of Machine Learning Research, 2010,11:2543−2596.

[5] Tao Q, Ma P, Zhang MH, *et al*. Individual convergence of stochastic optimization methods in machine learning. Journal of Data Acquisition and Processing, 2017,32(1):17−25 (in Chinese with English abstract).

[6] Tao Q, Gao QK, Jiang JY, Chu DJ. Survey of solving the optimization problems for sparse learning. Ruan Jian Xue Bao/Journal of Software, 2013,24(11):2498−2507 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4479.htm [doi: 10.3724/SP.J.1001.2013.04479]

[7] Schmidt M, Roux NL, Bach F. Convergence rates of inexact proximal-gradient methods for convex optimization. In: Advances in Neural Information Processing Systems. 2011. 1458−1466.

[8] Devolder O. Stochastic first order methods in smooth convex optimization. Université Catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2011.

[9] Nesterov Y. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Mathematics Doklady, 1983,27(2):372−376.

[10] Nemirovsky AS, Yudin DB. Problem Complexity and Method Efficiency in Optimization. Wiley-Interscience Series in Discrete Mathematics, John Wiley-Interscience Publication, 1983.

[11] Devolder O, Glineur F, Nesterov Y. First-order methods of smooth convex optimization with inexact oracle. Mathematical Programming, 2014,146(1−2):37−75.

[12] Honorio J. Convergence rates of biased stochastic optimization for learning sparse ising models. In: Proc. of the 29th Int'l Conf. on Machine Learning (ICML 2012). 2012. 257−264.

[13] Rasch J, Chambolle A. Inexact first-order primal-dual algorithms. arXiv Preprint arXiv:1803.10576, 2018.

[14] Shamir O, Zhang T. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In: Proc. of the 30th Int'l Conf. on Machine Learning (ICML 2013). 2013. 71−79.

[15] Tao W, Pan ZS, Chu DJ, Tao Q. The individual convergence of projected subgradient methods using the Nesterov's step-size strategy. Chinese Journal of Computers, 2018,1:164−176 (in Chinese with English abstract).

[16] Hazan E, Kale S. Projection-free online learning. In: Proc. of the 29th Int'l Conf. on Machine Learning (ICML 2012). 2012. 521−528.

[17] Mahdavi M, Yang T, Jin R, Zhu S, Yi J. Stochastic gradient descent with only one projection. In: Advances in Neural Information Processing Systems. 2012. 503−511.

[18] Liu J, Ye J. Efficient Euclidean projections in linear time. In: Proc. of the 26th Int'l Conf. on Machine Learning (ICML 2009). 2009. 657−664.

[19] Bertsekas DP, Nedić A, Ozdaglar AE. Convex Analysis and Optimization. Belmont: Athena Scientific, 2003.

[20] Tseng P. Approximation accuracy, gradient methods, and error bound for structured convex optimization. Mathematical Programming, 2010,125(2):263−295.

[21] Chen G, Teboulle M. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. SIAM Journal on Optimization, 1993,3(3):538−543. [doi: 10.1137/0803026]

[22] Rakhlin A, Shamir O, Sridharan K. Making gradient descent optimal for strongly convex stochastic optimization. In: Proc. of the 29th Int'l Conf. on Machine Learning (ICML 2012). 2012. 1571−1578.

[23] Shalev-Shwartz S, Singer Y, Srebro N. Pegasos: Primal estimated sub-gradient solver for SVM. In: Proc. of the 24th Int'l Conf. on Machine Learning (ICML 2007). 2007. 807−814.

[24] Liu J, Ji S, Ye J. SLEP: Sparse learning with efficient projections. Arizona State University, 2009,6(491):7.

[25] Duchi J, Shalev-Shwartz S, Singer Y. Efficient projections onto the l_1-ball for learning in high dimensions. In: Proc. of the 25th Int'l Conf. on Machine Learning (ICML 2008). 2008. 272−279.

[26] Tao W, Pan ZS, Chu DJ, Tao Q. The optimal individual convergence rate for the projected subgradient method with linear interpolation operation. Journal of Computer Research and Development, 2017,54(3):529−536 (in Chinese with English abstract).

[27] Bennett KP, Parrado-Hernández E. The interplay of optimization and machine learning research. Journal of Machine Learning Research, 2006,7:1265−1281.

[28] Qu Z, Peter R, Zhang T. Quartz: Randomized dual coordinate ascent with arbitrary sampling. In: Advances in Neural Information Processing Systems. 2015. 865−873.

附中文参考文献:

[5] 陶卿,马坡,张梦晗,等.机器学习随机优化方法的个体收敛性研究综述.数据采集与处理,2017,32(1):17−25.

[6] 陶卿,高乾坤,姜纪远,储德军.稀疏学习优化问题的求解综述.软件学报,2013,24(11):2498−2507. http://www.jos.org.cn/1000-9825/4479.htm [doi: 10.3724/SP.J.1001.2013.04479]

[15] 陶蔚,潘志松,储德军,陶卿.使用 Nesterov 步长策略投影次梯度方法的个体收敛性.计算机学报,2018,41(1):164−176.

[26] 陶蔚,潘志松,朱小辉,陶卿.线性插值投影次梯度方法的最优个体收敛速率.计算机研究与发展,2017,54(3):529−536.

附录 1. 定理 1 的证明

证明:结合凸函数的最优性条件及 hinge 损失函数的性质,有:

$$\begin{aligned}
&\eta_t(f(\boldsymbol{w}_{t+1})-f(\boldsymbol{y}))\\
&=\eta_t\big(f(\boldsymbol{y}_t)-f(\boldsymbol{y})\big)+\eta_t\big(f(\boldsymbol{w}_{t+1})-f(\boldsymbol{y}_t)\big)\\
&\leqslant \eta_t\langle \boldsymbol{g}(\boldsymbol{y}_t),\boldsymbol{y}_t-\boldsymbol{y}\rangle+\eta_t\langle \boldsymbol{g}(\boldsymbol{y}_t),\boldsymbol{w}_{t+1}-\boldsymbol{y}_t\rangle+\eta_t M\|\boldsymbol{w}_{t+1}-\boldsymbol{y}_t\|\\
&=\eta_t\langle \hat{\boldsymbol{g}}(\boldsymbol{y}_t),\boldsymbol{y}_t-\boldsymbol{y}\rangle+\eta_t\langle \hat{\boldsymbol{g}}(\boldsymbol{y}_t),\boldsymbol{w}_{t+1}-\boldsymbol{y}_t\rangle+\eta_t M\|\boldsymbol{w}_{t+1}-\boldsymbol{y}_t\|+\eta_t\langle \boldsymbol{\varepsilon}_t,\boldsymbol{w}_{t+1}-\boldsymbol{y}\rangle,
\end{aligned}$$

由引理 1 及算法迭代式子可得:

$$\eta_t\langle \hat{\boldsymbol{g}}(\boldsymbol{y}_t),\boldsymbol{w}_{t+1}-\boldsymbol{y}_t\rangle+\frac{1}{2}\|\boldsymbol{w}_{t+1}-\boldsymbol{y}_t\|^2+\frac{1}{2}\|\boldsymbol{w}_{t+1}-\boldsymbol{y}\|^2\leqslant\eta_t\langle \hat{\boldsymbol{g}}(\boldsymbol{y}_t),\boldsymbol{y}-\boldsymbol{y}_t\rangle+\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{y}_t\|^2,$$

则有:

$$\eta_t\left(f(\boldsymbol{w}_{t+1})-f(\boldsymbol{y})\right)\leqslant\eta_t\langle\hat{\boldsymbol{g}}(\boldsymbol{y}_t),\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\rangle+\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{y}_t\|^2-\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{w}_{t+1}\|^2-\frac{1}{2}\|\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\|^2+$$
$$\eta_t\langle\hat{\boldsymbol{g}}(\boldsymbol{y}_t),\boldsymbol{w}_{t+1}-\boldsymbol{y}_t\rangle+\eta_tM\|\boldsymbol{w}_{t+1}-\boldsymbol{y}_t\|+\eta_t\langle\boldsymbol{\varepsilon}_t,\boldsymbol{w}_{t+1}-\boldsymbol{y}\rangle,$$

其中,由 Young 不等式可得:

$$-\frac{1}{2}\|\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\|^2+\eta_tM\|\boldsymbol{w}_{t+1}-\boldsymbol{y}_t\|\leqslant\frac{\eta_t^2}{2}M^2,$$

$$\eta_t\left(f(\boldsymbol{w}_{t+1})-f(\boldsymbol{y})\right)\leqslant\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{y}_t\|^2-\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{w}_{t+1}\|^2+\frac{\eta_t^2}{2}M^2+\eta_t\langle\boldsymbol{\varepsilon}_t,\boldsymbol{w}_{t+1}-\boldsymbol{y}\rangle.$$

令 $\boldsymbol{y}=(1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{w}$, 代入可得:

$$\eta_t(f(\boldsymbol{w}_{t+1})-f((1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{w}))\leqslant\frac{1}{2}\|(1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{w}-\boldsymbol{y}_t\|^2-\frac{1}{2}\|(1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{w}-\boldsymbol{w}_{t+1}\|^2+$$
$$\frac{\eta_t^2}{2}M^2+\eta_t\langle\boldsymbol{\varepsilon}_t,\boldsymbol{w}_{t+1}-(1-\theta_t)\boldsymbol{w}_t-\theta_t\boldsymbol{w}\rangle.$$

令 $\boldsymbol{z}_t=-(\theta_t^{-1}-1)\boldsymbol{w}_t+\theta_t^{-1}\boldsymbol{y}_t$, 则有:

$$\frac{1}{2}\|(1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{w}-\boldsymbol{y}_t\|^2-\frac{1}{2}\|(1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{w}-\boldsymbol{w}_{t+1}\|^2=\frac{\theta_t^2}{2}\|\boldsymbol{w}-\boldsymbol{z}_t\|^2-\frac{\theta_t^2}{2}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|^2,$$

代入上面不等式,可得:

$$\eta_t(f(\boldsymbol{w}_{t+1})-f((1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{w}))\leqslant\frac{\theta_t^2}{2}\|\boldsymbol{w}-\boldsymbol{z}_t\|^2-\frac{\theta_t^2}{2}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|^2+\frac{\eta_t^2}{2}M^2+\eta_t\theta_t\langle\boldsymbol{\varepsilon}_t,\boldsymbol{w}-\boldsymbol{z}_{t+1}\rangle,$$

$$\eta_t(f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w}))\leqslant\eta_t(1-\theta_t)(f(\boldsymbol{w}_t)-f(\boldsymbol{w}))+\frac{\theta_t^2}{2}\|\boldsymbol{w}-\boldsymbol{z}_t\|^2-\frac{\theta_t^2}{2}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|^2+\frac{\eta_t^2}{2}M^2+\eta_t\theta_t\langle\boldsymbol{\varepsilon}_t,\boldsymbol{w}-\boldsymbol{z}_{t+1}\rangle,$$

$$\frac{1}{\theta_t^2}(f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w}))\leqslant\frac{1-\theta_t}{\theta_t^2}(f(\boldsymbol{w}_t)-f(\boldsymbol{w}))+\frac{1}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_t\|^2-\frac{1}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|^2+\frac{\eta_t}{2\theta_t^2}M^2+\eta_t\frac{1}{\theta_t}\langle\boldsymbol{\varepsilon}_t,\boldsymbol{w}-\boldsymbol{z}_{t+1}\rangle,$$

$$\frac{1-\theta_{t+1}}{\theta_{t+1}^2}(f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w}))\leqslant\frac{1-\theta_t}{\theta_t^2}(f(\boldsymbol{w}_t)-f(\boldsymbol{w}))+\frac{1}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_t\|^2-\frac{1}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|^2+\frac{\eta_t}{2\theta_t^2}M^2+\eta_t\frac{1}{\theta_t}\langle\boldsymbol{\varepsilon}_t,\boldsymbol{w}-\boldsymbol{z}_{t+1}\rangle,$$

递归可得:

$$\frac{1-\theta_{t+1}}{\theta_{t+1}^2}(f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w}))\leqslant\frac{1-\theta_0}{\theta_0^2}(f(\boldsymbol{w}_0)-f(\boldsymbol{w}))+\sum_{k=0}^{t}\frac{1}{2\eta_k}(\|\boldsymbol{w}-\boldsymbol{z}_k\|^2-\|\boldsymbol{w}-\boldsymbol{z}_{k+1}\|^2)+\sum_{k=0}^{t}\frac{\eta_k}{2\theta_k^2}M^2+\sum_{k=0}^{t}\frac{1}{\theta_k}\langle\boldsymbol{\varepsilon}_k,\boldsymbol{w}-\boldsymbol{z}_{k+1}\rangle,$$

由定义可知, $\boldsymbol{\varepsilon}$ 与 $\boldsymbol{w}$ 相互独立,结合引理 2 则有:

$$\sum_{k=0}^{t}\frac{1}{\theta_k}\langle\boldsymbol{\varepsilon}_k,\boldsymbol{w}-\boldsymbol{z}_{k+1}\rangle\leqslant\sum_{k=0}^{t}\frac{1}{\theta_k}\|\boldsymbol{\varepsilon}_k\|\|\boldsymbol{w}-\boldsymbol{z}_{k+1}\|\leqslant\sum_{k=0}^{t}\frac{1}{\theta_k}\|\boldsymbol{\varepsilon}_k\|D.$$

同时,由于 $\eta_k\leqslant\eta_{k-1}$,则有:

$$\sum_{k=0}^{t}\frac{1}{2\eta_k}(\|\boldsymbol{w}-\boldsymbol{z}_k\|^2-\|\boldsymbol{w}-\boldsymbol{z}_{k+1}\|^2)\leqslant\frac{1}{2\eta_0}\|\boldsymbol{w}-\boldsymbol{z}_0\|^2-\frac{1}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|^2+\sum_{k=0}^{t}\left(\frac{1}{2\eta_k}-\frac{1}{2\eta_{k-1}}\right)\|\boldsymbol{w}-\boldsymbol{z}_k\|^2\leqslant\frac{1}{2\eta_t}D^2,$$

结合以上不等式可得:

$$f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w})\leqslant\frac{\theta_t^2}{2\eta_t}D^2+\theta_t^2\sum_{k=0}^{t}\frac{\eta_k}{2\theta_k^2}M^2+\theta_t^2\sum_{k=0}^{t}\frac{1}{\theta_k}\|\boldsymbol{\varepsilon}_k\|D.$$

定理 1 得证. □

附录 2. 推论 1、推论 2 的证明

证明:结合定理 2,假设 $\boldsymbol{w}^*$为理论上的最优点,取步长 $\theta_k=\frac{1}{k+2}$,$\eta_k=\frac{1}{(k+2)\sqrt{k+2}}$, 则有:

$$E[f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w}^*)]\leqslant\frac{D^2+M^2}{2\sqrt{t+2}}+\sum_{k=0}^{t}\frac{k+2}{(t+2)^2}E[\|\boldsymbol{\varepsilon}_k\|]D.$$

根据 $E[\|\boldsymbol{\varepsilon}_k\|]$ 有界这个假设条件,可使得偏差项收敛为一个常数δ,则有:

$$E[f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w}^*)]\leqslant\frac{D^2+M^2}{2\sqrt{t+2}}+\delta.$$

推论 1 得证.推论 2 的证明类似于推论 1,略. □

刘宇翔(1992－),男,江西永丰人,硕士,主要研究领域为机器学习,模式识别.

程禹嘉(1996－),女,硕士,主要研究领域为机器学习,模式识别.

陶卿(1965－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为机器学习,模式识别,应用数学.